Resource-Efficient RGBD Aerial Tracking
–Supplementary Material–

Jinyu Yang[1,2,†], Shang Gao[1,†], Zhe Li[1,†], Feng Zheng[1,3*], Aleš Leonardis[2]
[1]Southern University of Science and Technology [2]University of Birmingham [3]Peng Cheng Laboratory
jinyu.yang96@outlook.com gaos2021@mail.sustech.edu.cn zhe.li.cs@outlook.com
f.zheng@ieee.org a.leonardis@cs.bham.ac.uk

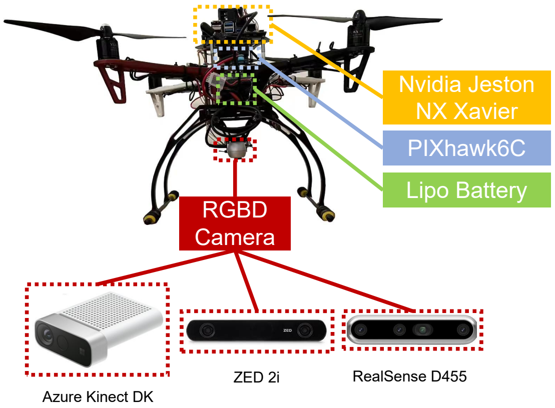

Figure 1. Overview of our data collection platform. Three alternatives are provided for capturing RGBD data. Note that RGBD cameras are connected with the drone by pan-tilt, thus the capturing viewpoints can be flexibly changed.

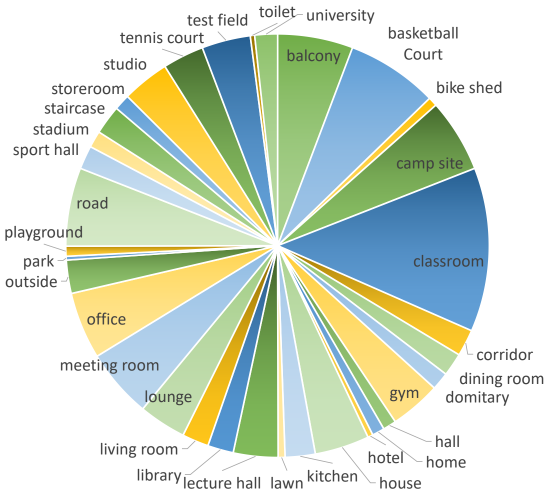

Figure 2. Data distribution of scenarios appeared in our test set.

A. Dataset Construction

Flight platforms. We present our real-world data collection on a handcrafted flight platform, mounted with advanced RGBD cameras, as shown in Fig. 1. There are three alternatives for RGBD video capturing, *i.e.*, Azure Kinect DK, ZED 2i and RealSense D455. They are used for video collection under different scenarios and different viewpoints, which can increase the dataset diversity on acquisition process. For ease of use, we also apply a compact commercial camera drone platform - *DJI Mavic Air 2* - to acquire high-quality RGB video streams, with which we then obtain depth maps by monocular depth estimation. This helps us to capture videos in some narrow spaces and guarantee the flight safety. To maintain high-quality depth information in the whole dataset, we employ DenseDepth [1] to generate corresponding depth maps.

Dataset statistics. We provide the distribution of captured scenarios in our test set in Fig. 2. As shown, 34 places are included in our test set, which covers diverse scenarios for generic aerial tracking evaluation. Specifically, our dataset includes indoor scenarios in human daily scenes, which provides potentials on broad applications of aerial robots. Besides, multiple viewpoints and challenges guarantee that the proposed D^{2}Cube maintains a high diversity.

B. More Results

Attribute-based performance. For in-depth analysis, we provide attribute-based performance in term of F-score on all compared trackers. The results are shown in Fig. 3. Obviously, ProTrack [6] shows outstanding performance on all attributes, while our proposed EMT ranks the second on 17 of 18 attributes with a very compact model size. Besides, most tested trackers show consistent trends in some attributes, *e.g.*, low performance on illumination variation and dark scenes, indicating that environment-level attributes are very challenging for state-of-the-art trackers. While trackers show much better performance on classical tracking challenges, *i.e.*, out-of-view, motion blur, scale variation and so on. In terms of RGBD trackers, DeT [5] performs well on scale variation, except for the outstanding performance of ProTrack and the proposed EMT. On the other hand, RGB trackers perform generally lower than RGBD trackers. Notably, some popular efficient trackers, *e.g.*, HiFT [2], DaSiamRPN [8] and TCTrack [3], show severe performance degradation in terms of illumination change, overexposure, and background clutter, demonstrating that current color-only aerial trackers are very sensitive to the overall appearance change.

Visualized results. To vividly show the performance of representative trackers on our proposed D^{2}Cube, we provide more visualized results in Fig. 4. The compared trackers include ProTrack [6], DeT [5], HCAT [4], UDAT [7] and the proposed EMT. 18 video sequences covering 18 attributes are shown for comparison. As shown, our EMT can perform well against most of the challenges. Specifically, our EMT can address difficulties like BC (background clutter) and CM (camera motion), in which tracking failures are presented by color-only trackers. Failed cases of EMT are given in OE (overexposure) and SF (sensor failure), which represent some extreme challenging tracking scenarios.

References

[1] I. Alhashim and P. Wonka. High quality monocular depth estimation via transfer learning. 2018. 1

[2] Ziang Cao, Changhong Fu, Junjie Ye, Bowen Li, and Yiming Li. Hift: Hierarchical feature transformer for aerial tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15457–15466, 2021. 2

[3] Ziang Cao, Ziyuan Huang, Liang Pan, Shiwei Zhang, Ziwei Liu, and Changhong Fu. Tctrack: Temporal contexts for aerial tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14798–14808, 2022. 2

[4] Xin Chen, Ben Kang, Dong Wang, Dongdong Li, and Huchuan Lu. Efficient visual tracking via hierarchical cross-attention transformer. *arXiv preprint arXiv:2203.13537*, 2022. 2

[5] Song Yan, Jinyu Yang, Jani Kapyla, Feng Zheng, Ales Leonardis, and Joni-Kristian Kamarainen. Depthtrack: Unveiling the power of rgbd tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10725–10733, 2021. 2

[6] Jinyu Yang, Zhe Li, Feng Zheng, Ales Leonardis, and Jingkuan Song. Prompting for multi-modal tracking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3492–3500, 2022. 1, 2

[7] Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. Unsupervised domain adaptation for nighttime aerial tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8896–8905, 2022. 2

[8] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018. 2

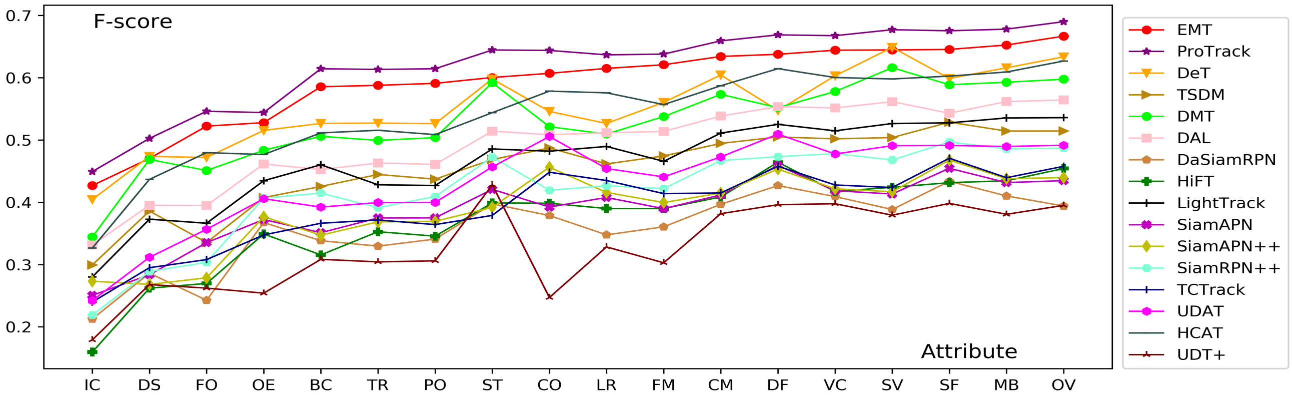

Figure 3. Attribute-based performance in terms of F-score. IC = Illumination Change, DS = Dark Scenes, FO = Full Occlusion, OE = Overexposure, BC = Background Clutter, TR = Target Rotation, PO = Partial Occlusion, ST = Similar Targets, CO = Composite Object, LR = Low Resolution, FM = Fast Motion, CM = Camera Motion, DF = Deformation, VC = Viewpoint Change, SV = Scale Variation, SF = Sensor Failure, MB = Motion Blur, OV = Out-of-view.

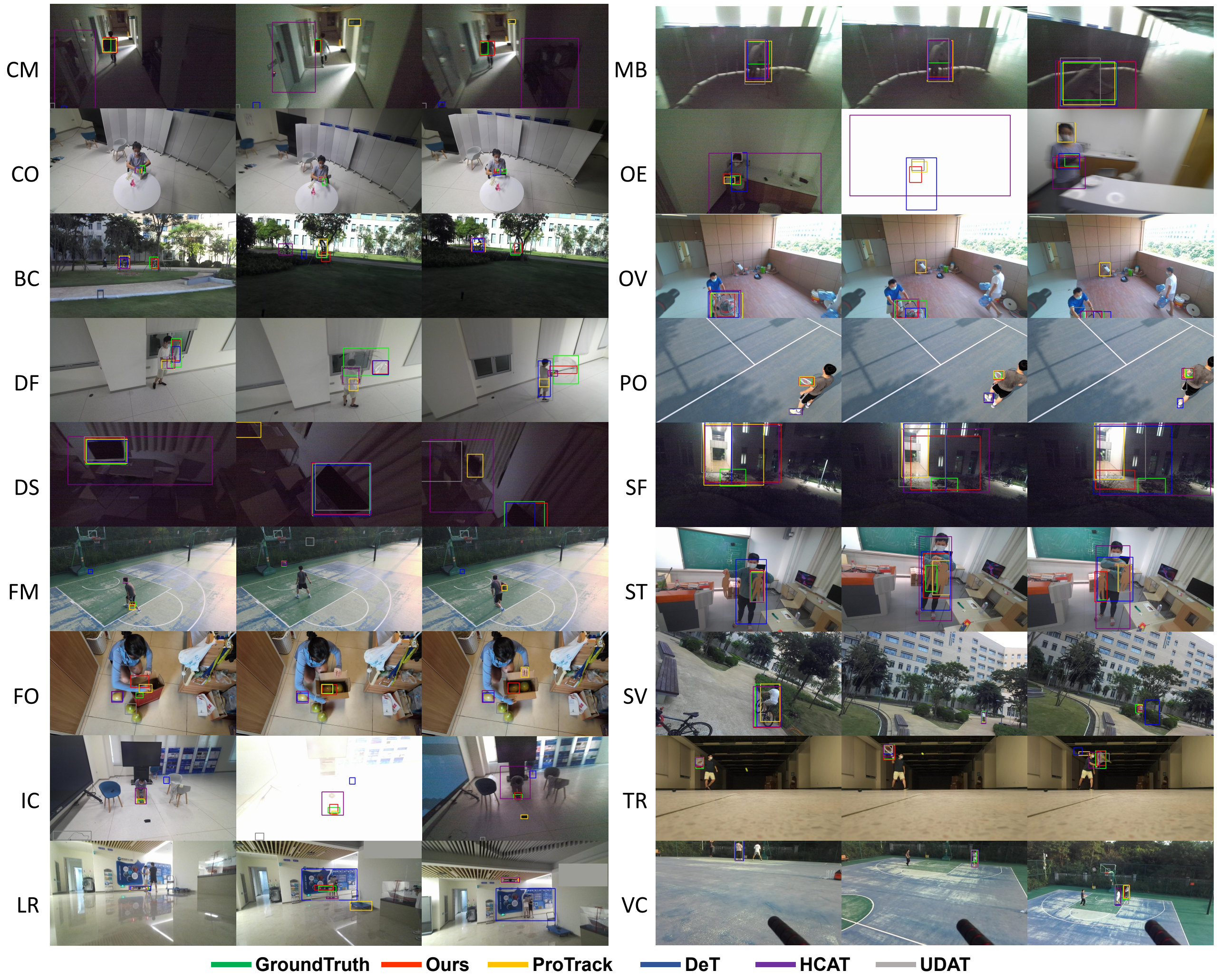

Figure 4. Visualized results for different challenges in D^{2}Cube. Zoom in for details.